基于 LDA 主题模型的用户电信轨迹恢复算法 *

徐广根[1]，杨　璐[1, 2]，严建峰[1, 2†]，徐彩旭[1]，石鸿斌[1]

(1. 苏州大学 计算机科学与技术学院，江苏 苏州 215006; 2. 香港城市大学 创意媒体学院，香港 中国)

摘　要：随着移动通信技术的发展和移动设备的普及，关于人们日常移动行为的轨迹数据记录愈发的丰富起来。海量的轨迹数据背后隐藏着关于人及人类社会的有价值的知识模式。为了使基于轨迹数据产生的知识模式更精准有效服务用户，能够准确、可靠地恢复缺失电信轨迹显得尤为重要。目前大多数方法主要针对 GPS 轨迹等连续轨迹进行建模，而缺乏对移动通信场景中产生的电信轨迹恢复的研究。因此，针对电信轨迹缺失恢复问题，将电信轨迹恢复问题转化为矩阵补全问题，提出了一种基于 LDA 主题模型的恢复算法。实验中，与传统矩阵补全算法进行综合比较，并观察了不同参数对轨迹恢复效果的影响。实验结果表明，与传统矩阵补全算法相比，运用 LDA 主题模型能够显著提高缺失电信轨迹的恢复精度。

关键词：电信轨迹；轨迹恢复；LDA 主题模型

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2018.01.0066

User telco trajectory recovery algorithm based on LDA topic model

Xu Guanggen[1], yang Lu[1, 2], Yan Jianfeng[1, 2†], Xu Caixu[1], Shi Hongbin[1]

(1. *School of Computer Science & Technology*, *Soochow University*, *Suzhou Jiangsu* 215006, *China*; 2. *School of Creative Media*, *City University of Hong Kong China*)

Abstract: With the development of mobile communication technology and the popularization of mobile devices, the daily track record data become rich. Massive track data hides valuable knowledge about person and human society. In order to make the knowledge model generated based on the trajectory data more accurate and effective to serve the users, it is particularly important to be able to recover the missing telco trajectories accurately and reliably. Currently, most of the methods mainly focus on modeling continuous trajectories such as GPS trajectories, but lack of researches on the restoration of telco trajectories generated in mobile communication scenarios. Therefore, it have transformed the problem of telecommunication trajectory recovery into a matrix completion problem, and proposed a recovery algorithm based on the LDA topic model. In the experiment, it make a comprehensive comparison with the traditional matrix completion algorithm and observe the effect of different parameters on trajectory recovery. The experimental results show that compared with the traditional matrix completion algorithm, the LDA topic model can significantly improve the recovery accuracy of missing telco tracks.

Key words: telco trajectory; trajectory recovery; LDA topic model

0　引言

在当前大数据时代，数据所带来的影响远远超出了企业领域，其不仅能带来商业价值，也能产生社会价值。随着信息通讯技术的发展，手机普及率越来越高，移动通信基站的覆盖率也越来越高。因此，研究手机定位技术获取的轨迹信息，不但可以为城市交通规划和出行方式的划分提供更有效决策[1, 2]，甚至可以利用节假日中用户的出行定位数据估计出可能会产生的客流高峰和区域，从而及时做好相应的交通管理。因此，如何精准地恢复用户轨迹的缺失部分，从而挖掘更丰富的知识模式，为用户提供人性化的服务具有极高的应用价值。移动对象轨迹缺失恢复研究引起了学术界的广泛关注。Lou 等人[3]提出了一种对于行驶过程中采样率很低的稀疏轨迹进行直接匹配的算法研究。但局限是忽略了时间间隔较大的稀疏轨迹。在时间间隔较大的情况下，该算法不能把匹配的路段拼接起来形成完整轨迹。Bernstein 等人[4]使用 GPS 轨迹数据统计得到车辆运行速度、

收稿日期：2018-01-22；**修回日期：**2018-03-21　**基金项目：**国家自然科学基金资助项目（61373092，61033013，61272449，61202029）；江苏省教育厅重大项目（12KJA520004）；江苏省科技支撑计划重点资助项目（BE2014005）

作者简介：徐广根（1992-），男，硕士研究生，主要研究方向为移动对象数据挖掘、机器学习、深度学习；杨璐（1982-），女，副教授，博士，主要研究方向为软件可靠性、机器学习；严建峰（1978-），男（通信作者），副教授，博士，主要研究方向为并行计算、机器学习（yanjf@suda.edu.cn）；徐彩旭（1993-），男，硕士研究生，主要研究方向为社交关系数据挖掘；石鸿斌（1992-），男，硕士研究生，主要研究方向为机器学习、数据挖掘.

方向拐角、车辆运行的连续性等特征，加上车辆前一段时段轨迹数据点匹配来分析采样率较低的轨迹数据点，根据车辆行驶的历史轨迹信息提高了稀疏轨迹点匹配的精度。Zheng 等人[5]根据车辆行驶的历史轨迹数据建立了一套路径推理系统，提高了路径推理的准确性。该算法可以快捷地匹配采样率低的稀疏轨迹，有效地恢复了采样率较低的稀疏轨迹缺失部分。罗宇等人[6,7]提出了一种基于隐马尔科夫模型的局部最优状态路径的轨迹恢复算法，利用最大后验概率构建轨迹的缺失状态。徐超等人[8]提出了还原个人出行 GPS 轨迹缺失的算法，通过设置经验阈值分析轨迹中缺失的情况，然后定义了最优路径的概念来搜索 GPS 轨迹缺失处的轨道站点，最终选择最优路径来还原 GPS 轨迹中的缺失路段。上述所有方案都为本文的研究提供了指导性作用，但存在的局限都是基于全球定位系统(global positioning system, GPS)技术[9]产生的轨迹恢复缺失时段的位置，而且都需要使用获取成本较高的路网数据。基于 GPS 定位的轨迹精度高、实时性高，而电信基站定位产生的轨迹数据量大、覆盖面积广，但精度低、离散程度高。针对上述方案的不足，本文充分考虑电信轨迹缺失的问题，对上海某运营商产生的用户电信轨迹在时空上完成预处理。具体过程是：在时间上将一天划分成若干个时间段，空间上城市基于 K-Means 聚类算法划分，再构建时空单词矩阵，将轨迹缺失恢复问题转换为矩阵缺失补全问题，最后运用 LDA 主题模型恢复缺失电信轨迹。

1　潜在狄里克雷分配

1.1　LDA 模型原理

LDA 模型是 Blei 等人[10]在 2003 年提出的一种对离散数据集（如文档集）建模的概率主题模型，是一种对文本数据的主题信息进行建模的方法，通过发现文档中抽象的主题，挖掘语义背后隐含的信息，从而有助于高效地处理大规模的文档集。LDA 模型是一个包含单词、主题、文档的三层贝叶斯网络概率图模型[11]。基于这样一种前提假设：文档是由若干个隐含主题构成，这些主题是由文本中若干个特定有效词汇构成，忽略文档中的句法结构和单词出现的先后顺序[12]。LDA 模型文档-主题-单词拓扑结构如图 1 所示。

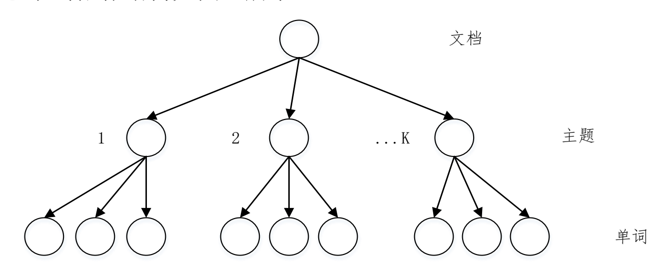


图 1　LDA 模型文档—主题—单词拓扑结构

图 2 是常用的 LDA 图模型的表示。在图 2 中圆圈表示变量；深色表示可观测变量；浅色表示超参数或隐含变量；方框表示重复循环；箭头表示依存关系。符号标签含义如表 1 所示。

表 1　符号标签含义

$1 \le d \le D$	文档索引
$1 \le n \le V$	单词索引
$1 \le k \le K$	主题索引
$\vec{\vartheta_d}$	文档 d 的主题多项分布（K 维向量）
$\vec{\varphi_k}$	主题 k 的单词多项分布（V 维向量）
$\Theta_{D\times K}$	文档-主题分布矩阵
$\Phi_{K\times V}$	主题-单词分布矩阵
N_d	文档 d 的长度
$z_{d,n}$	文档 d 中第 n 个单词的主题
$w_{d,n}$	文档 d 中第 n 个单词的词项
$\mathbf{z}_{D\times V} = \{z_{d,n}^k\}$	单词的主题标签
$\mathbf{w}_{D\times V} = \{w_{d,n}\}$	文档-单词矩阵
α	文档-主题分布的超参数
β	主题-单词分布的超参数

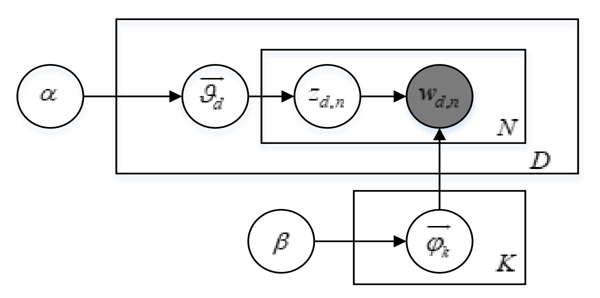


图 2　LDA 贝叶斯图模型

算法 1　LDA 概率主题模型生成文档集过程

//主题
1. for 所有的主题 $k \in [1, K]$:
2. 　　采样 $\vec{\varphi_k} \sim Dirichlet(\beta)$
3. end for
//文档
4. for 所有的文档 $d \in [1, D]$:
5. 　　采样 $N_d \sim Poisson(\zeta)$
6. 　　采样 $\vec{\vartheta_d} \sim Dirichlet(\alpha)$
7. 　　for 对文档 d 中所有的单词 $n \in [1, N_d]$:
8. 　　　　采样 $z_{d,n} \sim Multi(\vec{\vartheta_d})$
9. 　　　　采样 $w_{d,n} \sim Multi(\vec{\varphi_{z_{d,n}}})$
10. 　　end for
11. end for

根据生成文本过程，在给定先验参数 α 和 β 的条件下，可以得到文档集的联合概率分布（包括所有可观测变量和隐含变量）：

$$p(\mathbf{w},\mathbf{z},\Theta,\Phi \mid \alpha,\beta) = \prod_{k=1}^{K} p(\vec{\varphi_k} \mid \beta) \cdot \prod_{d=1}^{D} p(\vec{\vartheta_d} \mid \alpha) \cdot \prod_{n=1}^{N_d} p(z_{d,n} \mid \vec{\vartheta_d}) p(w_{d,n} \mid \vec{\varphi_{z_{d,n}}}) \quad (1)$$

对于生成单词 $w_{d,n} = t$ 的概率如下：

$$p(w_{d,n}=t\,|\,\vec{\vartheta}_d,\Phi)=\sum_{k=1}^{K}p(w_{d,n}=t\,|\,\vec{\varphi}_k)p(z_{d,n}=k\,|\,\vec{\vartheta}_d) \tag{2}$$

给定文档集后，LDA 模型的目标就是使得隐含变量 Z 的后验概率最大化（maximum a posteriori, MAP）。根据上述文档集生成过程，可以得到包含隐含变量的后验概率分布是

$$p(\mathbf{z},\Theta,\Phi\,|\,\mathbf{w},\alpha,\beta)=\frac{p(\mathbf{w},\mathbf{z},\Theta,\Phi\,|\,\alpha,\beta)}{p(\mathbf{w}\,|\,\alpha,\beta)} \tag{3}$$

其中：$p(\mathbf{w}\,|\,\alpha,\beta)$ 是文档集联合概率 $p(\mathbf{w},\mathbf{z},\Theta,\Phi\,|\,\alpha,\beta)$ 在文档主题分布 Θ 和主题单词分布 Φ 上的积分，计算复杂度高，所以需要可接受的近似推理算法来计算该后验概率。

1.2 Gibbs 采样

Gibbs 采样通过对难以求解的隐含变量的后验概率进行采样，迭代更新来获得 LDA 主题模型最终的参数。Gibbs 采样[13]基础理论是马尔可夫链蒙特卡洛[14]（Markov chain Monte Carlo, MCMC）。马尔可夫链的数学定义为

$$p(X_{t+1}=x\,|\,X_t,X_{t-1},\ldots)=p(X_{t+1}=x\,|\,X_t) \tag{4}$$

Gibbs 采样是相对简单的算法，经常用于近似推理高维度模型。LDA 主题模型的 Gibbs 采样算法把主题看做隐含变量，通过对文档集的联合分布中 Θ 和 Φ 进行积分消除，因为它们可以通过已经观测的单词 $w_{d,n}$ 和对应的主题标签 $z_{d,n}$ 联合统计得到。每个单词的主题标签 $z_{d,n}$ 是马尔可夫链上的状态变量，然后利用 MCMC 采样算法进行推理。这种通过一些参数进行积分的模型推理算法被称为"塌陷"，通过积分得到主题的后验分布为 $p(\mathbf{z}\,|\,\mathbf{w})$。

$$p(\mathbf{z}\,|\,\mathbf{w})=\frac{p(\mathbf{z},\mathbf{w})}{p(\mathbf{w})} \tag{5}$$

从式(5)可以得到，使用 Gibbs 采样算法，则需要计算主题和单词的联合概率分布。在 LDA 中，主题和单词的联合概率分布可以分解为 $p(\mathbf{z},\mathbf{w})=p(\mathbf{z}\,|\,\alpha)\cdot p(\mathbf{w}\,|\,\mathbf{z},\beta)$，这两个因子分别对应着 LDA 生成文档集的两个过程：

a) $\alpha\rightarrow\vec{\vartheta}_d\rightarrow z_{d,n}$ 表示生成文档中所有单词的主题。$\alpha\rightarrow\vec{\vartheta}_d$ 是 Dirichlet 分布，$\vec{\vartheta}_d\rightarrow z_{d,n}$ 是多项分布，所以整个过程是 Dirichlet-Multinomial 共轭结构。

$$\begin{aligned}p(\mathbf{z}\,|\,\alpha)&=\int p(\mathbf{z}\,|\,\Theta)p(\Theta\,|\,\alpha)d\Theta\\&=\int\prod_{d=1}^{D}\frac{1}{\Delta(\alpha)}\prod_{k=1}^{K}\vartheta_{d,k}^{n_d^{(k)}+\alpha_k-1}d\vec{\vartheta}_d\\&=\prod_{d=1}^{D}\frac{\Delta(\vec{n}_d+\alpha)}{\Delta(\alpha)},\quad\vec{n}_d=\{n_d^{(k)}\}_{k=1}^{K}\end{aligned} \tag{6}$$

其中：

$$\Delta(\alpha)=\frac{\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K}\alpha_k)} \tag{7}$$

$n_d^{(k)}$ 是文档 d 中主题 k 出现的次数。

b) $\beta\rightarrow\vec{\varphi}_k\rightarrow w_{d,n}$ 表示生成文档中所有单词。

$\beta\rightarrow\vec{\varphi}_k$ 是 Dirichlet 分布，$\vec{\varphi}_k\rightarrow w_{d,n}$ 是多项分布。用类似于 a)的推导方法可以得到

$$p(\mathbf{w}\,|\,\mathbf{z},\beta)=\prod_{k=1}^{K}\frac{\Delta(\vec{n}_k+\beta)}{\Delta(\beta)},\quad\vec{n}_k=\{n_k^{v}\}_{v=1}^{V} \tag{8}$$

其中：$n_k^{(v)}$ 是主题 k 中单词 v 出现的次数。

综合 a)和 b)，得到联合分布：

$$p(\mathbf{w},\mathbf{z}\,|\,\alpha,\beta)=\prod_{d=1}^{D}\frac{\Delta(\vec{n}_d+\alpha)}{\Delta(\alpha)}\cdot\prod_{k=1}^{K}\frac{\Delta(\vec{n}_k+\beta)}{\Delta(\beta)} \tag{9}$$

根据联合分布，求解下标 $i=(d,n)$，即第 d 篇文档中第 n 个单词的条件概率。令 $\vec{w}=\{w_i=t,\vec{w}_{\neg i}\}$，$\vec{z}=\{z_i=k,\vec{z}_{\neg i}\}$，则

$$\begin{aligned}p(z_i=k\,|\,\vec{z}_{\neg i},\vec{w})&=\frac{p(\vec{w},\vec{z})}{p(\vec{w},\vec{z}_{\neg i})}\\&=\frac{p(\vec{w},\vec{z})}{p(\vec{w}_{\neg i}\,|\,\vec{z}_{\neg i})p(w_i)}\cdot\frac{p(\vec{z})}{p(\vec{z}_{\neg i})}\\&\propto\frac{\Delta(\vec{n}_z+\beta)}{\Delta(\vec{n}_{z,\neg i}+\beta)}\cdot\frac{\Delta(\vec{n}_d+\alpha)}{\Delta(\vec{n}_{d,\neg i}+\alpha)}\\&=\frac{n_{k,\neg i}^{(t)}+\beta_t}{\sum_{t=1}^{V}n_{k,\neg i}^{(t)}+\beta_t}\cdot\frac{n_{d,\neg i}^{(k)}+\alpha_k}{(\sum_{k=1}^{K}n_d^{(k)}+\alpha_k)-1}\\&\propto\frac{n_{k,\neg i}^{(t)}+\beta_t}{\sum_{t=1}^{V}n_{k,\neg i}^{(t)}+\beta_t}(n_{d,\neg i}^{(k)}+\alpha_k)\end{aligned} \tag{10}$$

其中：$\vec{n}_d$ 是文档 d 的主题数向量；$\vec{n}_k$ 是主题 k 的单词数向量。因此，可以得到

$$\vartheta_{d,k}=\frac{n_d^{(k)}+\alpha_k}{\sum_{k=1}^{K}n_d^{(k)}+\alpha_k} \tag{11}$$

$$\varphi_{k,t}=\frac{n_k^{(t)}+\beta_t}{\sum_{t=1}^{V}n_k^{(t)}+\beta_t} \tag{12}$$

最终，Gibbs 采样算法完整训练流程如下：

算法 2 LDA 主题模型 Gibbs 采样算法训练过程

输入：文档集 $\mathbf{w}$，超参 α 和 β，主题数 K。

输出：Θ,Φ。

1. 随机为每篇文档的每个单词分配主题，根据这些主题初始化 $n_d^{(k)}$ 和 $n_k^{(t)}$，n_d 和 n_k
2. Repeat:
3. for 所有文档 $d\in[1,D]$:
4. for 文档 d 中的所有单词 $n\in[1,N_d]$:

//删除该单词的主题计数

5. $n_d^{(k)}--=1; n_d--=1; n_k^{(t)}--=1; n_k--=1;$

//根据式(10)采样得到该单词的新主题

6. $k \sim p(z_i \mid \vec{z}_{\neg i}, \vec{w})$

//增加该单词的主题计数

7. $n_d^{(k)}++=1; n_d++=1; n_k^{(t)}++=1; n_k++=1;$

8. end for

9. end for

10. Until 收敛或者达到设定的迭代次数

11. 根据式(11)和(12)计算得到 Θ, Φ

2　电信轨迹缺失恢复过程

与基于 GPS 定位的轨迹相比，更为丰富的电信轨迹来自于电信基站定位数据，定位精度较低。因此，恢复电信轨迹在某个具体时刻缺失的精确位置相对比较困难。为了简化恢复电信轨迹缺失的问题，本文将恢复电信轨迹在缺失时段所经过的某个区域。

2.1　构造时空单词矩阵

基于电信基站的定位任意用户每天的电信轨迹表示为 $Trj=\{(t_1,Tower_1),(t_2,Tower_2),\ldots,(t_n,Tower_n)\}$，其中：$t_i$ 表示用户接听通话、发送/接受短信、流量上网连接通信基站的时刻；$Tower_i$ 表示用户的移动设备在 t_i 时刻连接的基站位置，即 $Tower_i=(lon_i,lat_i)$；lon 和 lat 分别表示该基站的经纬度。

a) 时间分段：将一天 24 h 均匀划分成 T 个时段。若 $T=24$，则 0:00 与 1:00 之间的任意时刻都属于同一时段，以此类推。

b) 空间分块：根据运营商在城市各个区域部署的全部基站位置信息，利用 K-Means 算法[15]对基站位置进行聚类，从而将城市划分成 C 块区域。

完成时间分段和空间分块两个步骤后，可以将用户原始电信轨迹转换为 $Trj^{'}=\{(T_1,R_1),(T_2,R_2),\ldots,(T_n,R_n)\}$，从而最终生成时空单词矩阵 X，如图 3 所示。

	时段 1，区域 1	时段 1，区域2	⋮	时段 1，区域 C	时段 2，区域 1	⋮	时段 2，区域 C	⋮	时段 T，区域 C
用户 1，day1									
用户 1，day2									
…									
用户 1，dayM									
用户 2，day1									
…									
用户 2，dayM									
…									
用户 U，dayM									

图 3　时空单词矩阵 X

矩阵 X 的行表示所有用户在所有天生成的电信轨迹，行数等于用户数 U 与天数 M 的乘积；列表示所有时段所有区域构成的时空单词，列数等于时段数 T 与区域数 C 的乘积；值表示用户在指定时空单词上的访问次数，即用户在该时段内访问该区域内的所有电信基站次数。

2.2　划分训练集和测试集

本文将用户电信轨迹看做是记录用户一天出行的完整轨迹。为了模拟缺失电信轨迹恢复，对每条轨迹都需要随机地挖去一个时段内经过所有区域。如图 4 所示，阴影部分表示每条电信轨迹挖去时段的所有区域，即需要恢复该时段内的轨迹信息，作为测试集记为 $TestX$；剩余的白色部分作为训练集，残缺矩阵记为 $TrainX$。

	时段 1，区域 1	时段 1，区域2	⋮	时段 1，区域 C	时段 2，区域 1	⋮	时段 2，区域 C	⋮	时段 T，区域 C
用户 1，day1									
用户 1，day2									
…									
用户 1，dayM									
用户 2，day1									
…									
用户 2，dayM									
…									
用户 U，dayM									

图 4　处理后的时空单词矩阵 X

2.3　基于 LDA 主题模型的电信轨迹恢复算法

基于 LDA 主题模型的电信轨迹恢复算法主要思路是：将 2.2 节中得到的残缺矩阵 $TrainX$ 看做文档集，每条用户电信轨迹看做一篇文档，每个时段区域组合看做一个单词，用户在该时段区域组合内出现的次数作为该单词在对应文档中的词频，从而在该文档集上利用 LDA 主题模型算法得到文档—主题矩阵 Θ 和主题—单词矩阵 Φ。然后每条用户电信轨迹中被挖去时段内 C 块区域的权重可以通过 Θ 和 Φ 对应的行列权重向量相乘得到。最终选取权重最大的 N 块区域作为该缺失时段内可能经过的区域。具体流程如下：

算法 3　基于 LDA 主题模型的电信轨迹恢复方法

输入：　3.2 节中得到的 $TrainX$ 和 $TestX$，主题数 K，选取权重最大的区域数量 N。

输出：　每条用户电信轨迹恢复的 N 块区域。

1. 针对文档—单词矩阵 $TrainX$ 使用 LDA 主题模型算法得到文档—主题矩阵 Θ 和主题–单词矩阵 Φ。
2. 利用 Θ 和 Φ 对应的行列权重向量相乘计算得到 $TestX$ 中每条电信轨迹被挖去时段内 C 块区域权重。
3. 选取每条电信轨迹被挖去时段内权重最大的 N 块区域作为该缺失时段内可能经过的区域。

3　实验结果和分析

为了分析基于 LDA 主题模型对电信轨迹恢复方法的准确性和时间效率，对此进行了实验。实验硬件环境配置是：3 台 Huawei RH 2288 服务器组成的集群上，每台机器的配置都相同。CPU 为 Intel Xeon CPU E5-2690 v2@3.00 GHz，40 核，内存为 140 GB。实验抽样采集上海市某运营商 2016 年 7 月 12 号到 8 月 12 号的 10 万用户 CDR (call detail records) [16]数据、短信和上网记录构成的电信轨迹。抽样条件是：将每天 24 个小时均匀划分成 24 个时段，用户每天至少有 10 个不同时段存在基站连接记录。

3.1　评价标准

本文采用准确率(precision)、召回率(recall)和 F1-score 来衡量电信轨迹恢复的效果。当 N 越大，准确率会越低，召回率会越高。准确率定义为

$$Precision@N = \frac{|Y' \cap Y|}{N}$$

召回率定义为

$$Recall@N = \frac{|Y' \cap Y|}{|Y|}$$

F1-score 综合权衡了准确率和召回率，计算公式为

$$F1\text{-}score@N = 2 \times \frac{Recall@N \times Precision@N}{Recall@N + Precision@N}$$

其中：Y 表示当前电信轨迹在被挖去时段真实经过的区域集合；Y' 表示预测的 Top N 块区域集合。

3.2　不同方法比较

本文将电信轨迹恢复问题转换成矩阵缺失补全问题，本节将 LDA 主题模型算法和常见的传统矩阵补全算法非负矩阵分解 (Non-negative Matrix Factorization, NMF) [17]和概率矩阵分解 (Probabilistic Matrix Factorization, PMF) [18]从电信轨迹恢复精度和时间效率两个角度上完成对比分析。实验中设置时段数 $T = 24$，区域数 $C = 300$，主题数 $K \in \{10, 30, 50, 100, 150\}$，实验结果如表 2～6 和图 5 所示。

表 2　主题数 K=10

评价指标 \ N		@1	@2	@3	@4	@5
Precision	NMF	0.246	0.215	0.194	0.163	0.145
	PMF	0.273	0.256	0.224	0.196	0.178
	LDA	**0.354**	**0.297**	**0.267**	**0.246**	**0.231**
Recall	NMF	0.262	0.323	0.405	0.478	0.544
	PMF	0.281	0.368	0.453	0.516	0.586
	LDA	**0.318**	**0.444**	**0.534**	**0.603**	**0.660**
F1-score	NMF	0.254	0.258	0.262	0.243	0.229
	PMF	0.277	0.302	0.300	0.284	0.273
	LDA	**0.335**	**0.356**	**0.356**	**0.350**	**0.342**

表 3　主题数 K=30

评价指标 \ N		@1	@2	@3	@4	@5
Precision	NMF	0.252	0.223	0.208	0.185	0.157
	PMF	0.291	0.274	0.255	0.224	0.207
	LDA	**0.383**	**0.319**	**0.281**	**0.254**	**0.236**
Recall	NMF	0.271	0.335	0.413	0.493	0.562
	PMF	0.293	0.381	0.463	0.529	0.625
	LDA	**0.355**	**0.483**	**0.571**	**0.632**	**0.682**
F1-score	NMF	0.261	0.268	0.277	0.269	0.245
	PMF	0.292	0.319	0.329	0.315	0.311
	LDA	**0.369**	**0.384**	**0.376**	**0.363**	**0.350**

表 4　主题数 K=50

评价指标 \ N		@1	@2	@3	@4	@5
Precision	NMF	0.317	0.292	0.265	0.249	0.231
	PMF	0.348	0.313	0.287	0.267	0.244
	LDA	**0.414**	**0.354**	**0.338**	**0.294**	**0.266**
Recall	NMF	0.295	0.363	0.465	0.532	0.613
	PMF	0.314	0.418	0.492	0.557	0.656
	LDA	**0.387**	**0.524**	**0.593**	**0.644**	**0.708**
F1-score	NMF	0.306	0.324	0.338	0.339	0.336
	PMF	0.330	0.358	0.363	0.361	0.356
	LDA	**0.400**	**0.423**	**0.430**	**0.404**	**0.387**

表 5　主题数 K=100

评价指标 \ N		@1	@2	@3	@4	@5
Precision	NMF	0.357	0.324	0.303	0.296	0.275
	PMF	0.379	0.358	0.329	0.317	0.298
	LDA	**0.513**	**0.447**	**0.382**	**0.316**	**0.283**
Recall	NMF	0.336	0.391	0.477	0.549	0.633
	PMF	0.353	0.439	0.518	0.604	0.636
	LDA	**0.452**	**0.648**	**0.753**	**0.815**	**0.847**
F1-score	NMF	0.346	0.354	0.371	0.385	0.383
	PMF	0.366	0.394	0.402	0.416	0.406
	LDA	**0.481**	**0.529**	**0.507**	**0.455**	**0.424**

表 6　主题数 K=150

评价指标 \ N		@1	@2	@3	@4	@5
Precision	NMF	0.374	0.338	0.315	0.303	0.283
	PMF	0.412	0.376	0.355	0.346	0.327
	LDA	**0.493**	**0.426**	**0.366**	**0.304**	**0.276**
Recall	NMF	0.347	0.413	0.498	0.564	0.657
	PMF	0.368	0.448	0.532	0.588	0.608
	LDA	**0.431**	**0.616**	**0.728**	**0.783**	**0.833**
F1-score	NMF	0.360	0.372	0.386	0.394	0.396
	PMF	0.389	0.409	0.426	0.436	0.425
	LDA	**0.460**	**0.504**	**0.487**	**0.438**	**0.415**

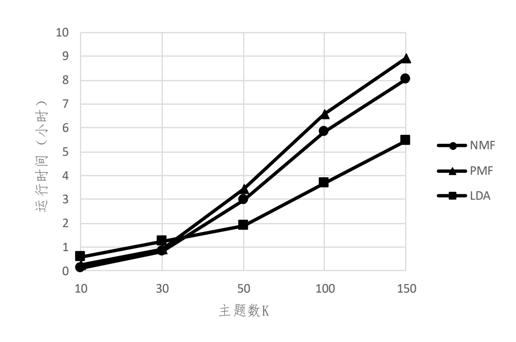


图 5　各种算法运行时间对比

从表 2～6 可以看出，随着主题数 K 增加，NMF 和 PMF 在 precision、recall 和 F1-score 都会逐渐增加，PMF 轨迹恢复效果略优于 NMF；随着主题数 K 增加，LDA 在 precision、recall 和 F1-score 是先增加后有略微减小，当 $K=100$ 时，precision、recall 和 F1-score 达到最高，且明显高于 NMF 和 PMF。从图 5 中可以看出，当主题数 K 较小时，LDA 运行时间会略高于 NMF 和 PMF；随着主题数 K 增加，LDA 运行时间逐渐少于 NMF 和 PMF。主要原因是：NMF 和 PMF 的时间复杂度是关于主题数 K 的多项式级别，而 LDA 的时间复杂度是关于主题数 K 的线性级别。因此，可以得出结论：基于 LDA 主题模型的用户电信轨迹恢复精度和时间效率都优于 NMF 和 PMF 两种算法，充分表明将用户电信轨迹类比成文档可以有效地描述用户每天出行的行为模式。因此，相比传统矩阵缺失补全算法，使用 LDA 主题模型恢复时空矩阵中的缺失部分更具优势。

从上述实验结果中可以看出，选择 LDA 主题模型且设置主题数 $K=100$ 在电信轨迹恢复的精度和时间效率上会取得最佳效果。在此基础上，根据实验结果可以计算比较一周内周一到周日和一天内各个不同时段的所有电信轨迹的平均恢复精度。为简化对比结果，只选取 precision@1、recall@1、F1-score@1 作为评价标准。实验结果分别如图 6 和 7 所示。

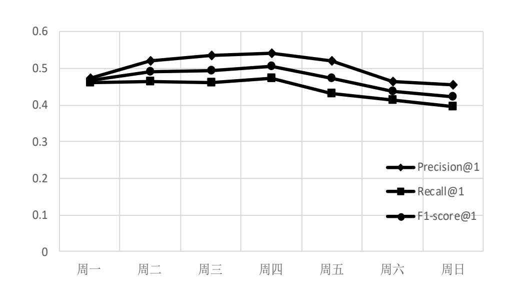


图 6　一周内各天电信轨迹平均恢复精度对比

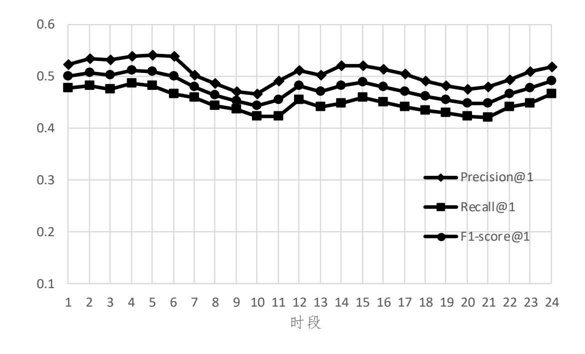


图 7　一天内各个时段电信轨迹平均恢复精度对比

从图 6 中可以看出，电信轨迹恢复的精度在周一到周五高于周六和周日，周二到周四较稳定。实验设置时段数 $T=24$，则每个时段长度是 1 h。从图 7 中可以看出，电信轨迹恢复的精度在凌晨和夜间高于白天，上午 9 点电信轨迹恢复精度最低。产生上述现象的主要原因是：用户工作日的出行轨迹相比较周末更有规律；用户在凌晨和夜晚大多数处于休息状态，在白天出行活动范围更大。因此，图 6 和 7 进一步验证了 Song 等人[19]提出的人类的行为有其潜在的规律性。

4　结束语

本文提出了一种基于 LDA 主题模型恢复用户电信轨迹的方法。在此过程中介绍了 LDA 主题模型原理和 Gibbs 采样推理算法；然后本文详细地阐述了利用 LDA 主题模型解决电信轨迹恢复问题方法和步骤；最终本文通过实验对比了 NMF、PMF、LDA 主题模型三种不同算法的恢复精度和时间效率，综合得出结论 LDA 主题模型可以有效地解决用户电信轨迹恢复问题。

本文中研究的不足之处是仅仅能恢复电信轨迹在某个时段属于某个区域，并不能恢复电信轨迹某个时刻的具体位置，所以电信轨迹恢复还需要进一步研究。在未来发展中，可以考虑把用户个人兴趣爱好、路况、天气等上下文信息加入到模型中，进一步提高恢复的精度。

参考文献：

[1] Zhang Zelun, Stefan P. A new post correction algorithm (PoCoA) for improved transportation mode recognition [C]// Proc of IEEE International Conference on Systems, Man, and Cybernetics. 2013: 1512-1518.

[2] Sasank R, Mun M Y, Burke J, *et al.* Using mobile phones to determine transportation modes [J]. ACM Trans on Sensor Networks, 2010, 6 (2): 13.

[3] Lou Yin, Zhang Chengyang, Zheng Yu, *et al.* Map-matching for low-sampling-rate GPS trajectories [C]// Proc of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information systems. 2009: 352-361.

[4] Bernstein D, Kornhauser A. An introduction to map matching for personal navigation assistants [J]. Geometric Distributions, 1998, 122 (7): 1082-1083.

[5] Zheng Kai, Zheng Yu, Xie Xing, *et al.* Reducing uncertainty of low-sampling-rate trajectories [C]// Proc of the 28th IEEE International Conference on Data Engineering. 2012: 1144-1155.

[6] 罗宇，杜利民. 基于隐马尔可夫模型局部最优状态路径的数据重建算法 [J]. 电子与信息学报, 2004, 26 (5): 722-726. (Luo Yu, Du Limin. HMM local optimal state path-based data imputation [J]. Journal of Electronics & Information Technology, 2004, 26 (5): 722-726.)

[7] 苏腾荣，吴及，王作英，等. 利用空间相关性的改进 HMM 模型 [J]. 计算机工程与设计, 2010, 31 (5): 1023-1026. (Su Tengrong, Wu Ji, Wang Zuoying, *et al.* Improved HMM model using spatial correlation [J]. Computer Engineering and Design, 2010, 31 (5): 1023-1026.)

[8] 徐超，季民河．个人出行轨迹中轨道交通段 GPS 信号缺失修补算法 [J]. 交通信息与安全, 2012, 30 (4): 6-10. (Xu Chao, Ji Minhe. An algorithm for repairing missing trajectories of GPS data in personal light-rail travel [J]. Computer and Communications, 2012, 30 (4): 6-10.)

[9] Hofmann-Wellenhof B, Lichtenegger H, Collins J. Global positioning system: theory and practice [M]. [S. l.] : Springer Science & Business Media, 2012.

[10] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of machine Learning research, 2003, 3 (1): 993-1022.

[11] Nir F, Dan G, Moises G. Bayesian network classifiers [J]. Machine Learning, 1997, 29 (2-3): 131-163.

[12] 姚全珠，宋志理，彭程．基于 LDA 模型的文本分类研究 [J]. 计算机工程与应用, 2011, 47 (13): 150-153. (Yao Quanzhu, Song Zhili, Peng Cheng. Research on text categorization based on LDA [J]. Computer Engineering and Applications, 2011, 47 (13): 150-153.)

[13] Ludden T M, Mu Song. Markov chain Monte Carlo and gibbs sampling [Z]. 2004.

[14] George E I, McCulloch R E. Variable selection via Gibbs sampling [J]. Journal of the American Statistical Association, 1993, 88 (423): 881-889.

[15] Burkardt J. K-means clustering [C]// Advanced Research Computing, Interdisciplinary Center for Applied Mathematics. 2009.

[16] Perkins I F C. System and method for processing call detail records: U. S. Patent 6, 658, 099 [P]. 2003-12-2.

[17] Farial S, Berry M W, Paul P V, *et al.* Document clustering using nonnegative matrix factorization [J]. Information Processing & Management, 2006, 42 (2): 373-386.

[18] Andriy M, Ruslan S. Probabilistic matrix factorization [C]// Advances in Neural Information Processing Systems. 2008: 1257-1264.

[19] Song Chaoming, Qu Zehui, Blumm N, Albert-László Barabási. Limits of predictability in human mobility [J]. Science, 2010, 327 (5968): 1018-1021.